Automated Knowledge Extraction from Internet for a Crisis Communication Portal

Ong Sing Goh and Chun Che Fung

School of Information Technology, Murdoch University,
Murdoch, Western Australia, 6150, Australia
os.goh@murdoch.edu.au, l.fung@murdoch.edu.au

Abstract. This paper describes the development of an Automated Knowledge Extraction Agent (AKEA) which was designed to acquire online news and document from the internet for the establishment of a knowledge based crisis communication portal. It was recognized that in times of crisis, an effective communication mechanism is essential to maintain peace and calmness in the community by providing timely and appropriate information. It is proposed that the incorporation of software agents into the crisis communication portal will be capable to send alert news to subscribed users via internet and mobile services. The proposed system consists of crawler, wrapper, name-entity tagger, AIML (Artificial Intelligence Markup language) and an animated character is used in the front-end for human computer communication.

1 Introduction

With the acceptance and increasingly reliance of the Internet, the Internet has now become "the" repository of human knowledge and information for the 21st century. On the other hand, advancements in internet and mobile communication technologies have provided effective and cheap means of communication for the modern society. The global implications of such technologies are unparalleled in the history of human civilization. Hence, the Internet now serves two of the most important functions in the modern world – as a giant virtual storehouse of data, information and knowledge, and, as the true information superhighway whereby delivery of all kinds of data and information can be done cheaply and quickly.

The potential of effective use of these two aspects are particularly important in times of crisis. Within the context of this paper, crisis may be referred to events or incidents that have the potential to cause national panic, confusion, unrest and possible catastrophe. These crises may be due to health epidemic, natural disasters and man-made tragedies such as terrorist attacks. Examples of these events that happened in the recent past are Severe Acute Respiratory Syndrome (SARS), bird flu, mad cow disease, September 11, earth quakes and tsunami. In these cases, accurate information delivered within the shortest duration of time at the lowest costs would be essential in informing the affected communities and the relevant authorities. In particular, if decisions are made quickly and appropriately, this will have the benefits of reducing the potential damages and will lead to better manage of the situations.

L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 1226 – 1235, 2005.

This paper reports the development an Automated Knowledge Extraction Agent (AKEA) which was designed to establish the knowledge base for a global crisis communication system called CCNet. CCNet was proposed during the height of the SARS epidemic in 2003. It was aimed at providing up-to-date information to its users via a conversational software robot called AINI (Artificial Intelligence Neural-network Identity). The purpose of AINI is to deliver essential information from trusted sources and is able to interact with its users by animated characters. The idea is to rely on a human-like communication approach thereby providing a sense of comfort and familiarity. The functionalities of AINI have been reported in the past and development on AINI is ongoing [1]. It is foreseeable that the combination of AINI and AKEA will produce a more natural means of communication and computing in the near future.

1.1 Objectives of the Research

The objectives of this research are:

a. To develop a global crisis communication portal (CCNet portal) in order to provide the latest information for public awareness on the knowledge about and how to respond to a particular crisis.
b. To establish a new and effective human-computer communication approach by transforming traditional websites with static text and images to "humanized" websites by deploying Artificial Intelligent Neural-network Identity (AINI).
c. To develop an Automated Knowledge Extraction Agent (AKEA) to automatically build and enhance the knowledge base for the AINI conversation software robot.
d. To develop new approaches in internet and communication technologies for the effective distribution of information to the global communities.

2 Architectural Overview

The architectural design of the proposed system is shown in Figure 1. The CCNet Portal can be divided into two main parts plus a middle-tier of multiple knowledge bases. The two main parts are termed the *Front-End*, responsible for interaction with the user, and, the *Back-End*, which is designed to establish the knowledge bases in the background.

The AINI Server and Mobile Gateway are located in the middle. They function as the interconnection linkage between the Front-End (Client) and the Back-End (Server) of the system. They process the communication between the users of the system and the CCNet Portal. The AINI's engine comprises of an intelligent agent framework. All communications with AINI are carried out through a natural human-machine interface that uses natural language processing and speech technologies via a 3D animated character. AINI's engine carries out the sophisticated decision making process based on the information it interprets from the knowledge bases. These decision-making capabilities are based on the knowledge embedded in the XML specifications. The input and output of the modules in the AINI knowledge bases such as Expression Emotion, Customers and AlertNews are stored in XML-encoded data

structure. These modules are representations of the knowledge conceptualized in the format of XML data structure. From the perspective of the users, the CCNet system accepts questions and requests, and it is also capable to process the queries based on the information contained in AINI's knowledge base.

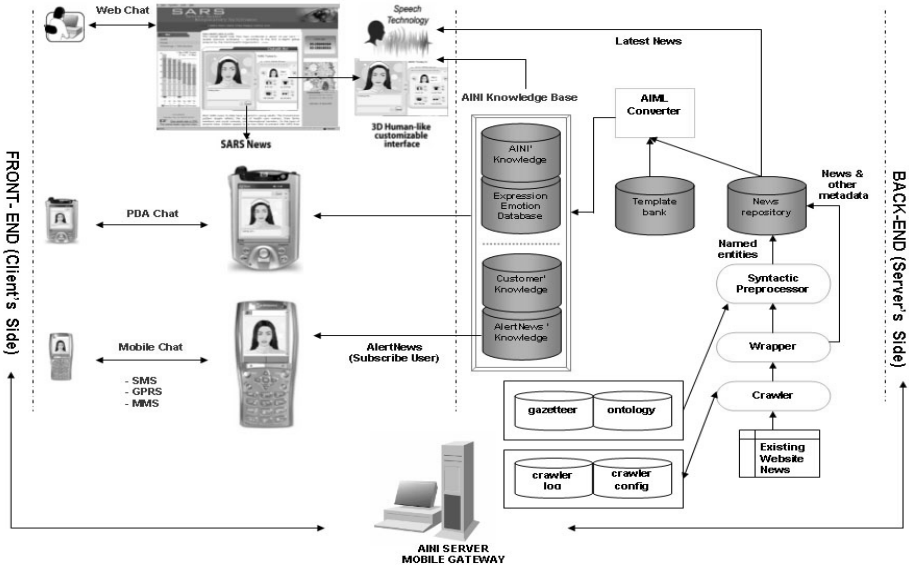

Fig. 1. Architecture Overview of CCNet

2.1 Front-End (Client's Side)

The Front-end provides the necessary interaction between the user and the system. Three different modes of communication are provided - web chat, PDA chat and mobile chat.

Web Chat

The web chat sessions allow interaction between a user and the software robot. The communication can text-based or voice-based with the animation of a 3D character. If voice is desired, Text-to-Speech technology is used to convert the text to voice using synthesizer hardware and software. This is particularly useful for someone who has difficulties or unfamiliar with the conventional keyboard. In terms of the animated character, users may customize the interface as required. They can also input the questions and receive the responses directly from the website. In addition, users may navigate through all the information on the topics or issues of their interest. If necessary, guidance may also be provided to assist the users.

AINI's Processing Engine

The main objective of AINI is to intelligently offer related information on various topics in a virtual environment where no real live agents or specialists are required to be physically involved. AINI uses natural language parsing in Artificial Intelligent

Mark-up Language (AIML) to search the predefined knowledge base as well as other data sources located in other systems via the internet. Users interact with the virtual advisor through WebGuide, WebTips and WebSearch engines. WebGuide is used to guide users through the entire portal. The WebTips engine, on the other hand, provides tips or hints to the users. The WebSearch system is an integrated search engine which can search for local sites as well as the Internet and online databases.

At the same time, the users can interact and chat with the AINI chatterbot or Virtual Agent. The chatterbot is based on natural-language processing and aimed to initiate conversations with users [1]. On the other hand, AINI also offers messaging, email and phone services to the users.

PDA Chat

Developing AINI into Personal Digital Assistance (PDA) devices is a recent approach in order to provide an alternative human and personalized interface between the computer and human. The PDA chat has the same functions as in web chats but with mobile capability. It is designed to incorporate mobile technology with natural language interface to assist interaction naturally with mobile devices. Implementation of PDA chat with the knowledge base was designed using WiFi technology.

2.2 Back-End System (Server's Side)

In this paper, the focus is on the development of a knowledge base which forms the "brain" of the CCNet portal. This knowledge base contains the domain knowledge for crisis communication based on specific discipline or topic. All the information in this knowledge base is going to be extracted from AKEA, which is explained in detail in the next section.

3 Establishment of AINI'S Knowledge Bases

In this proposed system, AINI's knowledge base consists of a common knowledge base, an expression emotion database, a customer knowledge base and an Alert-News knowledge base. From literature, it was identified that START (SynTactic Analysis using Reversible Transformations) developed by Boris Katz at MIT's Artificial Intelligence Laboratory is a natural language understanding system, and Omnibase is a virtual database that provides uniform access to heterogeneous and distributed Web sources via a wrapper-based framework [2]. A simplified version of the natural language annotation technology is employed here as the database access schemata to mediate between natural language and database queries. A detailed description of each component is provided in the following sections.

3.1 Common Knowledge Base

AIML is used to represent AINI's common knowledge base. It is an XML specification for programming chat robots created by ALICE Artificial Intelligence Foundation. A typical way of representing knowledge in an AIML file is as follows:

```
<aiml>
     <category>
             <pattern>PATTERN</pattern>
             <template>TEMPLATE</template>
     </category>
</aiml>
```

The <*aiml*> tag demonstrates that this file describes the way that knowledge is stored. The <*category*> tag indicates an AIML category and it is the basic unit of the chatterbot's knowledge. Each category has a <*pattern*> and a corresponding <*template*>. This <*pattern*> represents the question and the <*template*> represents the answer [3]. A user chats with AINI in the cyberspace and the topic may involve any topic related to crisis communication.

3.2 Expression Emotion Knowledge Base

The Expression Emotion Database, on the other hand, is used to identify and classify emotions within the context of the conversation between the user and the software robot. AINI was designed to perceive the emotion behind the human's input and it generates appropriate responses. The concept of communication between a human and the agent through AINI is depicted in Figure 2.

Human speech is passed to the natural processing unit in AINI for analysis and processing as shown in Figure 2. The natural processing unit generates proper responses by extracting the knowledge stored in the database. The emotion recognition unit is responsible for identifying the emotion found in the speech and instructs the agent to display an appropriate facial expression. For example, the agent will display a happy face to greet the user when the conversation starts; it will display a sad face when it hears something miserable; and it will show an angry face when the user says some obscene words.

An <*agplay/*> tag or "agent play" tag is created to produce an attractive expression for the character of that animated agent. Below is an example showing how the <*agplay/*> tag is used. This category is executed when the user greets the AINI by typing "HELLO". In return, the AINI will smile to the user and respond by saying "Hi there! How do you feel today?"

```
<category>
      <pattern>HELLO</pattern>
         <template>
             Hi there! How do you feel today?
                 <agplay anims="greet, pleased"/>
         </template>
</category>
```

3.3 AlertNews and Customer Knowledge Base

The AlertNews knowledge base provides news and information to users who use mobile chat via SMS, MMS or GPRS technologies. There are three types of users of this system - subscribed users, non-subscribed users and the CCNet editorial. The AlertNews architecture is shown in Figure 3.

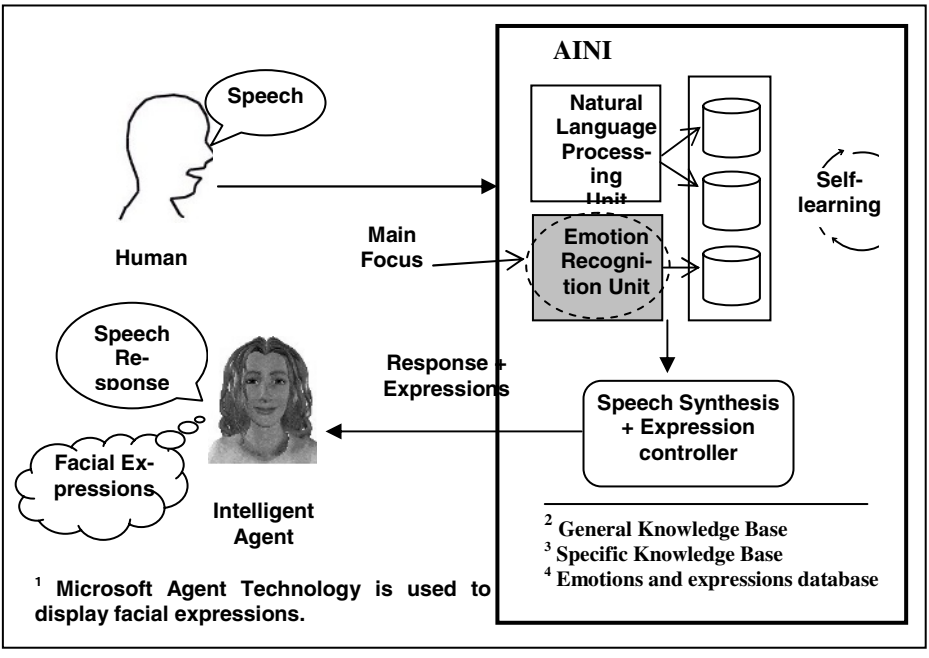

Fig. 2. Human-liked Emotion and Expression between user and AINI

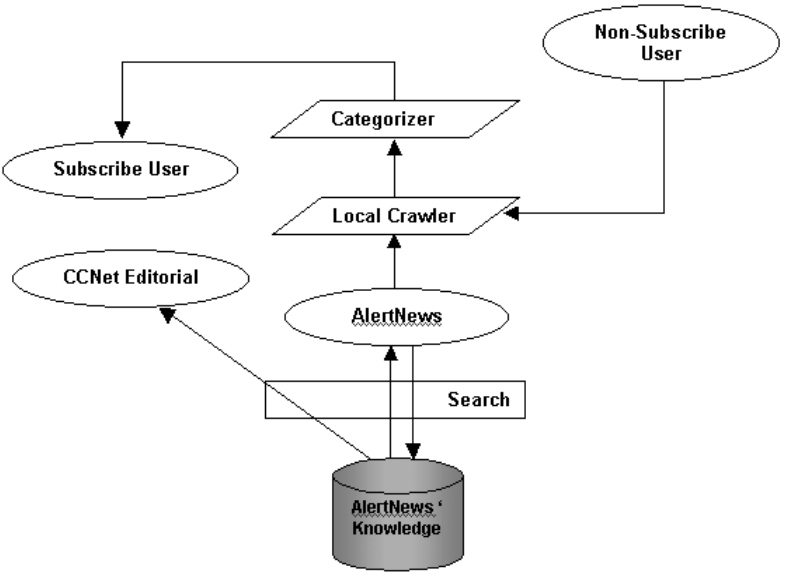

Fig. 3. AlertNews Architecture

Figure 4 shows the architecture of the CCNet Automated Knowledge extraction Agent (AKEA) focused on extracting information from the World Wide Web (WWW) for the AINI customer knowledge base. Four modules make up the agent. The modules are the Crawler, Wrapper, Named-entity Tagger and the AIML Converter. The crawler is the interface between the agent and the web. The functions of the crawler are like those used in conventional crawler-based search engines. The crawler resolves root domain names such as who.org, info.gov.hk, sars.gov.sg, etc. and follows subsequent links that are available on a page until a certain depth as defined by the user. These configurations are set in the crawler config database. For every page crawled, a copy is returned for further processing by the wrapper. The activities of the crawler are logged in the crawler log database.

The input to the syntactic preprocessor is the online news documents which may consist of several paragraphs. The functional model of the preprocessing phases required is shown in Figure 4 as part of the knowledge base construction system. Some of the stages in Figure 4 can be reorganised and even removed depending on the functions of the remaining phases. In other words, the choice of algorithm for a particular phase will determine the relevancy of other phases. This property is called *dependant optionality* where a particular phase is considered as redundant if the output it provides is already contained in the output of other phases. For example, the morphology analyzer can be put aside if the sentence parser also performs morphology analysis implicitly as part of its function.

Online news documents returned by the crawler are in the hypertext format and contain of a variety of unwanted characters. The Wrapper prepares the raw information by separating the actual news content and other meta-information from the hypertext characters or tags. This process is known as *cleaning* and the result is referred to as *cleaned news*. Once the information is processed, the key elements such as date of news, news title, news content and other relevant information are extracted and stored in the CCNet news repository.

The syntactic preprocessor performs the task of identifying the dependencies among the words. Based on the dependencies, grammatical relations (i.e. phrasal categories) like noun phrases, verb phrases and prepositional phrases are extracted using sentence parser for the English language like Link Grammar [4]. The named entities in noun phrases are assigned with tags such as disease, location and person using the weighted gazetteer approach proposed by Wong, Goh and et al. in [5]. A reference list in the Gazetteer is used by the preprocessor. These tags enable the agent to identify what type of entity the corresponding noun phrases are and in which level and node do these entities belong to in the ontology. Pronouns are also resolved whenever necessary. The named entities that have been tagged are inserted into the corresponding entry in the news repository. The syntactic preprocessor managed to identify two named entities namely meningitis and Burkina Faso. Using the gazetteer, the preprocessor will discover that meningitis is a type of disease and Burkina Faso is a country and tag them respectively using the ontology tag in the form of *named_entity[ontology tag]*.

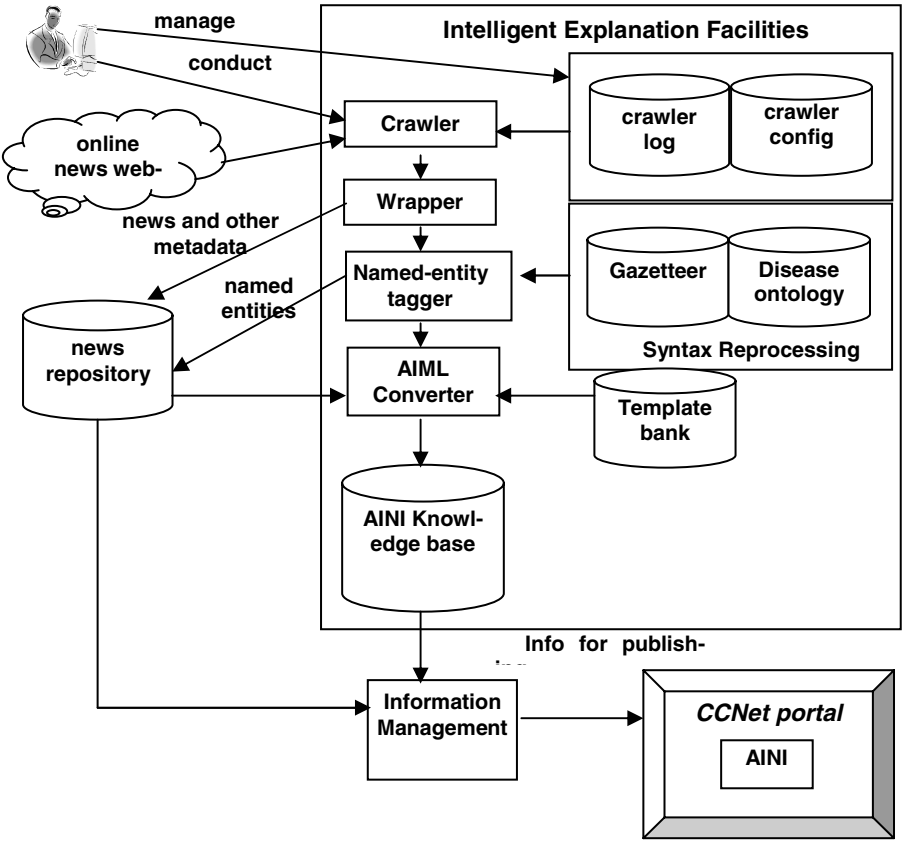

Fig. 4. CCNet Automated knowledge extraction agent architecture

In the gazetteer, each entry has additional information like *weight*, *ontology id* and the acceptable preceding/foregoing grammatical relations in addition to the triggering information, category and entity name. For example, a returned noun phrase "*Japanese Encephalitis disease"*, could trigger ambiguity. This could be resolved by just using the weighting mechanism without the need for any hand-crafted rules.

The information in the news repository is fed into two main components, namely the CCNet portal and the AIML converter. Information in the news repository is directly published to the CCNet portal without any further processing. The AIML converter uses the template bank to transform the news repository entries into AIML representation, which will be populated into AINI's knowledge base. The richer the template bank, the wider the scope of questions AINI will be able to handle. Substitutions will be made to the template using relevant values of each entry in the news repository. There are four fields in the template namely the *wh-token* corresponding to the ontology tag, first two lines of content, disease named entity and URL. The first and second require some processing prior to replacement.

The ontology tag associated with each named entity is resolved to obtain the corresponding wh-token. Currently, the agent is capable of handling four types of wh-token: *where*, resolved from *location* named-entities; *when*, resolved from *date* named-entities; *who*, resolved from *agent* named-entities and *what*. The *what* token is resolvable from all ontological entities with additional tokens. For example, given the named entity *Burkina Faso* and its tag *country*, we can obtain the *where* token and *what* token with the *country* tag. This is possible because the question *where does meningitis…*? is similar to asking *what country does meningitis…?*

The second processing required prior to replacement is to truncate the news content to the first two lines to be used in AINI's answers. The remaining news will be presented as part of a URL push. The AIML converter follows precedence in converting named-entities and their ontology tag into AIML representation. All questions handled by AINI are based on the concept of disease and thus, all news content will surely contain *disease* named-entities. During conversion, priority will be given to entities other than disease. Only when a news item that does not contain any other entities (i.e. there are no information about *location*, *person* or *date*) is encountered, then the converter will resolve the sole *disease* named-entity to the *what* token. Finally, these instances will be populated into AINI's knowledge base for learning and used by the AINI's chat interface in the CCNet portal for natural language question answering.

3.4 Multilevel Knowledge Base Natural Language Query

This section explains how AINI knowledge works. Firstly, AINI will search for an answer from Level 1 specific domain (Crisis Communication knowledge base) created by AKEA. If an answer is not found, it will move to Level 2 from the Frequently Asked Questions (FAQ) knowledge base which are stored in the FAQ table on MySQL database. Questions such as *what is SARS, how SARS spread*, *what is SARS vaccination*, *where SARS happen*, etc can be answered at this level. The next level is Level 3, which is metadata (News database). The search is done by identifying the keyword in the question and matching it with the content of the metadata. Since the WWW is so big that simple pattern matching techniques can often replace the need to understand both the structure and meaning of language. If an answer is still not found, AINI will proceed to Level 4 where AIML common knowledge will take place. If AINI still cannot answer the question, the last step will store the unanswered question in the database for the attention of the administrator. The answers will then be subsequently stored in the Level 2. This will enable AINI to answer the same question in the future.

4 Conclusion

The Internet has become a vital source of information and channel for effective communication during times of crisis and occurrence of global issues. This research will continue to develop and make use of the Internet to create global virtual communities by using intelligent agents and software robot. This intelligent software robot will assist the communication process by giving necessary and vital information needed by users during a crisis. It will also help in maintaining calmness and order so that the

country and the communities will not be hijacked by fear and panic. It is proposed that users will have more trust in the information provided by the intelligent software robot because of its interactive features and an ongoing engagement with the users. Furthermore, the integrated Text-To-Speech Technology and 3D human-like character or avatar in the system is capable to deliver speech and interact with the user in a humanlike manner thereby generating a sense of care and comfort. The portal also provides news, advertisements, conversation logs and statistics in the system benefiting the researchers in their efforts to further enhance the system. In the anticipated forthcoming epidemics or waves of new diseases due to mutation of bacteria and viruses, the output from this research will be useful to tackle future health crises. Information on natural disasters such as tsunamis, typhoons, earthquakes and floods will also be effectively managed and disseminated by deploying CCNet System. It is believed that CCNet provides a well-engineered platform for experimentation with various Web-enabled question answering techniques by employing conversation software robot. The implementation is currently under ongoing development. Progress and preliminary results will be reported. In the future, we will endeavour to continually refine the existing technology and to develop new frameworks in order to enable an efficient Internet-based global crisis communication.

Acknowledgements

This research is funded by KUTKM Grant under contract number PJP/2003/FTMK(1) (S017) and administered by the Centre for University Industry. The project is continued with a grant supported by the Murdoch University Division of Arts Research Excellence Grant Scheme in 2005.

References

1. Goh, O.S. and Fung C.C. (2003), "Intelligent Agent Technology in E-Commerce" in Jiming Liu, Yiuming Cheung, Hujun Yin (Eds.), *Intelligent Data Engineering and Automated Learning,* LNCS, Vol. 2690, Springer-Verlag, pp. 10-17.
2. Katz B, Felshin S, Yuret D, Ibrahim A, Lin J, Marton G, McFarland AJ, Temelkuran B (2002), "Omnibase: Uniform access to heterogeneous data for question answering", Natural Language Processing and language Processing and Information Systems. Lecture in Computer Science, Vol. 2553: pp. 230-234.
3. Goh, O. S. & Teoh, K. K. (2002), "Intelligent virtual doctor system" in Proceedings of the *2nd IEE Seminar on Appropriate Medical Technology for Developing Countries,* London, United Kingdom.
4. Lafferty, J.; Sleator, D.; Temperley, D. (1993), "Grammatical trigrams: a probabilistic model of link grammar", *Probabilistic Approaches to Natural Language. Papers from the 1992 AAAI Fall Symposium*, 1993, p 89-97.
5. Wong, W., Goh, O. S., & Mokhtar Mohd Yusof. 2004. "Syntax preprocessing in cyberlaw web knowledge base construction". In M. Mohammadian (Ed.), Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA 2004), Gold Coast, Australia. ISBN: 174-088-1893, pp: 174-184.